DOCUMENT RESUME

ED 085 660 CS 000 851

AUTHOR Pikulski, John J.
TITLE Criterion Referenced Measures for Clinical Evaluations.
PUB DATE Nov 73
NOTE 16p.; (Paper presented at the Annual Meeting of the College Reading Assn. (17th, Silver Springs, Md., Nov. 1-3, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; *Measurement Techniques; Norm Referenced Tests; Reading; Reading Ability; *Reading Clinics; Reading Diagnosis; Reading Improvement; Reading Instruction; *Reading Skills; *Reading Tests; Standardized Tests; Test Construction

ABSTRACT

This paper discusses criterion referenced tests' characteristics and their use in clinical evaluation. The distinction between diagnostic tests and criterion referenced measures is largely a matter of emphasis. Some authorities believe that in diagnostic testing the emphasis is upon an evaluation of an individual's strengths and weaknesses in skills areas with attention to the possible causes of problems that exist. Criterion referenced testing is less concerned with definition of disability and does not emphasize the etiology of problems. In constructing norm referenced test items, the writers of the test must select items of varying levels of difficulty. In criterion referenced measurement the scale is usually anchored at the extremes--one at the top indicating complete or perfect mastery of some defined abilities; one at the bottom indicating absence of some skills. The items on a criterion referenced measurement should be representative of skills that are essential to learning to read. The items should be arranged in an established hierarchy that would be used for teaching reading. In a clinical situation where the number of students is small, criterion referenced measures are simply ways of determining whether goals have been met. (WR)

CRITERION REFERENCED MEASURES FOR CLINICAL EVALUATIONS

John J. Pikulski

University of Delaware

Newark, Delaware

The term "criterion referenced testing" has been discussed at least since the early 1960's (Glaser, 1963), though some maintain that the concept goes back much further into educational testing history (See Ebel, 1971). It has had an opportunity to be called by many other names: domain referenced testing, edumetric testing, mastery tests, maximum performance tests, content referenced tests and undoubtedly many others as well. It has reached the point where there is disagreement as to how it should be defined, what form it should take and what its utility is. In short, the concept appears to have been through its initiation period with its accompanying debate and sensationalism. Reading specialists have become excited and confused about it, and it may now have reached the point where we can gain some perspective as to its importance to the field of reading and reading diagnosis. The general concept will be reviewed in the first part of this paper, and its possible implications for clinical evaluation will be discussed in the latter half.

Although there is not universal agreement about the definition of criterion referenced testing, there are some generally accepted parameters for the concept. The first step that is usually taken in defining it is to contrast it with norm referenced testing and sometimes with diagnostic testing. With norm referenced evaluation the basic question asks how well an individual performs when he is compared with other individuals. Under most circumstances the individual's performance is compared with the group of individuals who comprised the standardization population for the test that is being administered. As a result of the test the

examiner can draw comparisons with local, regional or national groups, depending on the population to which the individual is being compared. The focus throughout is upon inter-individual differences. The greater the variability among the individual performances, the better the test is evaluated in most circumstances.

The distinction between diagnostic tests and criterion referenced or norm referenced measures is largely a matter of emphasis, and many writers would prefer not to make a distinction between them. However, some feel that in diagnostic testing the emphasis is upon an evaluation of an individual's strengths and weaknesses in skill areas with attention to the possible causes of problems that exist. The primary purpose is to distinguish between individuals who do and do not have a learning deficiency. Criterion referenced testing is less concerned with definition of disability and does not emphasize the etiology of problems.

Prescott (1971) introduces another frequently cited consideration when he maintains that it is more accurate to discuss criterion-referenced interpretation of test scores rather than criterion referenced tests. By his distinction he raises another question. How different are norm referenced and criterion referenced measures in their make-up and in the types of items that they use? Proger and Mann (1973), Popham and Husek (1969), Simon (1969), and Prescott (1971) suggest that a test could be administered and then interpreted either from a norm or criterion referenced point of view. This position suggests that there are no identifying characteristics that separate norm referenced and criterion referenced tests apart from their interpretation. Some would hold that the only necessary characteristic of a criterion referenced test is that some standard be established for determining whether or not a particular skill has been achieved. In other words, the score necessary for "passing" must be set ahead of time without reference to the scores obtained by others taking the test. However, some writers maintain that other characteristics must be present as well.

The next few paragraphs will briefly discuss some of the characte 'stics that have been cited as necessary hallmarks of criterion referenced tests. Those listed below do not exhaust the possibilities.

Clear definition of task, objective, or skill that is to be measured. Millman (1972) maintains that there cannot be a criterion referenced test of general reading because an instructor would not be able to specify what a student could and could not do. Unfortunately, he does not elaborate upon this point, which is a controversial one. Many maintain that objectives must be stated in behavioral terms and tested in a fashion that allows for clear evidence as to whether a skill does or does not exist. For example, one published criterion referenced measure (Hackett, 1971, p. 17) sets the following objective: "Given a list of words the pupil identifies the three sounds of ea with 95% accuracy." A test item would be: "Choose the answer that tells how the letters ea sound in each word."

1. death

   A. long a

   B. long e

   C. short e

Supposedly, the student would need to be able to read the word and then choose one of the three alternatives. The item certainly satisfies the criterion of specificity, although it would raise other objections to be discussed below for some advocates of criterion referenced testing.

Would the following objective also satisfy the specificity requirement? The student will be able to read a newspaper. Carver (1972) cites this as an illustration of an important piece of information about a student's reading ability in his discussion of criterion referenced measurement. If one were to take a newspaper, select an article and ask a student to read it, would it be a criterion referenced

test? Is it specific enough or would it be a general test of reading? It would, of course, be necessary to define an acceptable level of performance ahead of time.

Items chosen for mastery rather than for obtaining a normal distribution. In constructing norm referenced test items, the writers of the test must select items of varying levels of difficulty. As Ebel (1971) points out, "a norm referenced measure is anchored in the middle, on an average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performances above and below the average level. In criterion referenced measurement the scale is usually anchored at the extremes--one at the top indicating complete or perfect mastery of some defined abilities; one at the bottom indicating complete absence of those abilities" (p. 282). In norm referenced measurement the goal for choosing items is to discriminate among individuals; in criterion referenced measures the objective is very different, the goal for each item is the same--determining whether a skill has or has not been mastered. Items are chosen because they are considered to be representative, essential skills needed by students, and the goal of the teacher is to have all students respond to all items with perfect or near perfect accuracy. The items must be carefully chosen. Mastery must be a realistic goal. Prescott (1971) also points out that mastery can be justified only "if the skill is demonstrably prerequisite to learning at a later stage" (p. 350). There is fairly good agreement on the part of test constructors that terms for norm referenced and criterion referenced tests are chosen in very different ways.

The tasks to be accomplished must be evaluated in the context of a normal reading situation. Johnson & Kress (1971) challenge the assumption that most tests that are norm referenced can be used for criterion referenced interpretation. They maintain that a multiple choice format is contradictory to the basic purpose for criterion referenced measures, which they maintain is finding out if the student can accomplish

the reading task. Returning to the illustration cited above (determining the sound of ea in death), these authors would most likely prefer to have the student read the word death as a way of determining his knowledge of the ea digraph rather than using the multiple choice format shown on page 3. Critics of this position would probably raise the objection that the student might be able to read the word death as a sight word and not know that the ea digraph has a short e sound. Quite likely those in harmony with the Johnson & Kress position would ask if he really needs to know that ea in death has a short e sound, so long as he can read the word correctly. Bleismer (1972) in advocating the greater use of more "natural" evaluation techniques raises the question of whether it isn't possible for a child to be able to accurately syllabicate a list of thirty words and yet not be able to pronounce any of them. He asks if it wouldn't be a more valid procedure to observe what the student does when he meets an unknown polysyllabic word in his reading material.

The items should be representative of skills that are essential to learning to read. This characteristic is of paramount importance to criterion referenced measurement in reading. The primary validity question is a content validity one. There is, of course, far less than perfect agreement as to what is and is not an essential skill for reading. More illustrations will be offered later in this paper. Since in criterion referenced testing the goal is mastery of a skill by all students, it is particularly wasteful in time and effort for groups of students to engage in achieving a minute goal that is not essential to some larger, useful goal. Critics of programs stressing hundreds of behavioral objectives for reading have charged that many of the specific objectives are not needed by many readers. Is it, for example, necessary to be able to apply phonic generalizations to nonsense words as is sometimes required? Is this not a highly artificial activity? When the child is asked to read a word, doesn't he use a combination of skills rather than a particular

one? If this is so, stressing ability to deal with a highly specific skill in isolation may be inappropriate.

The items should be arranged in an established hierarchy or should follow the sequence of a program that will be used for teaching reading.

Prescott (1971) is very firm regarding the first half of this characteristic of criterion referenced testing. He writes, "Basically, the criterion-referenced approach is of little value unless the assumption is made that mastery of one skill or bit of information is essential for mastery of another skill or bit of knowledge of a somewhat similar character at a higher level of difficulty or complexity" (p. 352). He goes on to acknowledge that there is no well accepted hierarchy of reading skills. Different programs follow different sequences. Many of the skills are not hierarchical in the sense of depending on lower-level skills. Some children may use context clues effectively before demonstrating a knowledge of phonic skills or vice versa. One could therefore conclude that using criterion referenced testing at this stage of knowledge regarding the reading process is inappropriate. The more reasonable approach, however, seems to be to arrange the criterion referenced testing of skills in the order in which they appear in a program by which a child is being instructed. This suggests that criterion referenced testing operates most efficiently when considered in relationship to some program. The program need not be a published one. It can be teacher devised, but it does require that the teacher designing the program have some suggested sequence of skill development in mind.

If skills are hierarchically arranged, there is much greater potential for determining necessary steps to overcome problems.

Criterion Referenced Testing Applied to Clinical Evaluation

Most of the journal articles which have discussed criterion referenced testing have appeared in the last three or four years. Most of them are unfortunately general

in their discussion, or else they seem devoted to defending some characteristic that the author feels is essential to criterion referenced testing. The remainder of this paper will be devoted to trying to consider the concept of criterion referenced testing to clinical evaluation in a somewhat directed way. However, much generality is unavoidable since clinic evaluations vary considerably from place to place.

The first prerequisite is, of course, that the clinic and its staff have a clear idea of the purpose of the evaluation. In general, there appear to be two general situations. In the first are clinics which have as their role a determination of the etiology of a reading problem and/or pattern of reading strength and weakness; no responsibility for treatment of the problem is assumed. The purpose is solely diagnosis and not remediation. With respect to the question of etiology, it may well be that there are greater opportunities for criterion referenced testing than is currently the case; however, discussion of criterion referenced measures other than reading is really outside the scope of this paper.

After etiological considerations, the diagnostic evaluation of reading probably calls for some determination of whether or not a problem in skill development exists. Norm referenced tests are probably necessary here. The diagnostician will probably use them, even if in a somewhat disguised form. For example, it is not logically or physiologically necessary to conclude that a child of average intelligence should begin to read at the end of a first grade experience. However, if such a child were evaluated and found to be unable to read, almost all clinics would conclude that he is a child with a reading problem.

Informal reading inventories are the main measurement instruments in some reading clinics, and many who support this form of evaluation point out that informal reading inventories are criterion referenced measures. However, it does seem that this is an instance where the interpretations and not the material make for the distinction between norm referenced and criterion referenced tests. If the diagnostician works

in a clinic and uses the same I.R.I. materials with all those tested, he must believe, for example, that most sixth grade youngsters are able to acceptably read a passage designated as being at sixth grade. How did it get that designation to begin with? Didn't the author of the material, who labeled it, have some basis for having decided that most sixth graders would be able to read it? His norming may be unsystematic, imprecise and implicit, but it's there, although perhaps in a rather crude form. If then the clinician pronounces the child sick or healthy in reading on the basis of his ability to deal with a particular selection, isn't he engaging in norm referencing evaluation? In no way is this to imply that this is an inappropriate procedure. The point is raised simply to suggest that in clinical testing we usually rely on some form of norm referencing in determining whether a problem does or does not exist. The question of the most valid materials and tests for drawing this conclusion is too broad a consideration to be dealt with here. Criterion referenced testing, however, is very limited in its ability to answer the question, "Is there or is there not a problem in reading?"

Measures such as informal reading inventories and some standardized measures as well might then be used by a diagnostic clinic for criterion referenced interpretation which in turn would lead to a remediation plan. Difficulties, however, begin to arise at this point if the diagnostician is not very familiar with the details of the treatment program to which the child might be exposed. As noted above, some maintain that criterion referenced measures should either follow an established hierarchy of skills known to be essential to reading or they should correspond closely to the program to be followed. Most diagnostic clinics deal with children who will be moving to schools or specific treatment facilities which do not have or would not use materials recommended by a diagnostic clinic. So long as this is the case, the use of criterion referenced testing by diagnostic clinics will remain limited.

Recommendations will necessarily be rather general since some skills may be essential and hierarchically important for one program and not for another. It does appear to this writer that the concept of criterion referenced testing can be fully used only if there is close correspondence between evaluation and instruction.

Clinical programs that provide instruction aimed at remediating reading problems have two additional goals beyond those dealt with by diagnostic clinics. They must obtain information regarding suitable reading materials and programs, and they have an obligation to evaluate progress that results from remediation intervention. It is in these two areas that criterion referenced measures are eminently helpful.

For some time we have been suggesting that we diagnose a youngster's needs and then decide on the type of program that would be of greatest benefit to him. This is probably the sequence that would be followed in the best of all possible reading worlds; however, in real life we probably have limited treatment alternatives and we match the child, as best we can, to one of these. If realistically this is the case, then the appropriate diagnostic procedure from a criterion referenced point of view would be to have samples of these materials available and to conduct a series of abbreviated teaching sessions to determine which techniques and materials are most beneficial for the child. The materials would have to be analyzed ahead of time for the skills required for dealing with them. Specific skills and goals would be focused upon, and the examiner could determine whether or not they had been mastered. This is the way in which informal reading inventories become criterion referenced. If the clinician accepts the assumption that a child will make fastest progress when working with materials that he can orally read with 95% accuracy, then what could be better than sampling from some potential instructional reading materials? This should not be the end, however, as it sometimes is. The material should lend itself to determining the existence or nonexistence of specific skills. In many cases, teacher guides or

notes will help define what these skills are. Child errors in reading this material and answering questions about it should provide the basis for determining whether or not these skills exist. Examiner-teachers should also be capable of analyzing errors of a general sort to set specific instructional goals. For example, a goal for a child who disregards punctuation when reading might be to have the child show a change in inflection and stop reading when he encounters a period.

In the past some reading clinic programs gave very little consideration to skill development with the emphasis being only on general reading ability. However, it is quite likely that much could be gained by working toward mastery of a specific skill, like being able to attach the appropriate sound value to the ch digraph. The criterion referenced testing then provides some structure and direction for the instructional program, and the instructional sessions in turn become the opportunity for continuous evaluation. The student in his reading will encounter other words which contain the ch digraph, and as he attempts to read them, the teacher can determine whether or not mastery has taken place for this specific skill. The above suggestion implies that the evaluation program fits into the instructional program rather than vice versa.

At the other end of the continuum from the remedial programs that dealt only with general reading are those that deal with hundreds of specific objectives. The sequence is: testing in order to determine if a skill can be demonstrated in isolation; teaching that skill; and testing to be sure it's attained. Methods and materials become subserviant to minute skills. Two dangers exist. Some students can successfully deal with items like the following illustration and yet may not be able to apply the skill when they encounter real reading material. The objective is: Given a list of words, the pupil identifies silent consonants with 95% accuracy. The directions are: Choose the word that has the silent consonant.

1. A. gnat
   B. gallant
   C. given

2. A. grind
   B. gnarl
   C. glory

(From Hackett, 1971, p. 21)

The other possible misuse of this approach is that students may be perfectly capable of applying the skill that is attempted to be measured in an item and yet not be able to complete the item successfully. For example: The objective is: "Given specified two syllable words and counterparts divided into syllables the pupil classified words by syllable stress with 95% accuracy." The directions are: Read each list of words. Locate the words that has the accent on the wrong syllable.

1. A. pe' ri od
   B. stran ger'
   C. sub scribe'

2. A. ex' act ly
   B. law' yer
   C. hel' met

(From Hackett, 1971, p. 34)

Ransom (1972) raises the question of whether the mastery of a succession of specific tests will make "adequate and eager readers" (p. 283). Considerations of the pros and cons of criterion referenced testing of very specific goals are very similar to considerations made regarding the pros and cons of the use of behavioral objectives

in teaching reading. Niles (1972) does an excellent job of summarizing the dangers and potential advantages of such an approach.

Ebel (1971) comes to the conclusion that criterion referenced measurements cannot "be expected to significantly improve our evaluations of educational achievements" (p. 287). One of his arguments is particularly relevant at this point. He maintains that the degree of detail in the specification of outcomes makes such an approach "unrealistic to expect and impractical to use ... the formulation of specific objectives which would suffice, costs more in time and effort than they are worth in most cases. Furthermore, if they are used, they are more likely to suppress rather than stimulate effective teaching" (p. 284). This may be a realistic conclusion. For example, Proger and Mann (1973) in advocating the clinical use of criterion referenced measurement, recommend that two or three full-time staff members spend an entire academic year in the specification of behavioral objectives and the writing of test items. This would result in objectives and tests for the first year students; additional planning time would be necessary for subsequent years. It's not difficult to understand how this plan would receive a very lukewarm acceptance by administrators.

The other major diagnostic obligation of a clinic that undertakes remediation is that of evaluating and gauging progress. Norm referenced tests may be particularly inappropriate here. Hieronymous (1972) points out the need for allowing different pupils in the same room to take several different levels of a test at the same time. This procedure, he feels, would use tests "that are more relevant, more accurate and less frustrating than tests intended for the 'average' pupils in the 'grade'" (p. 267). This frequently is not done. For example, in one remediation program that I recently reviewed the teachers were required to administer the age appropriate level of a widely used standardized test to their junior high students. The students took the same test at the conclusion of the year. The test was far too challenging even

for youngsters who make progress during the year. Their scores in both pre and post tests were essentially chance level scores on these tests which were designed for "average" readers. Some of these students had moved from being essentially non-readers to being able to deal first reader level materials--a highly significant breakthrough; yet none of this progress appeared in these norm referenced scores. In fact, some of these students, who actually had made progress, appeared to regress in skill according to the test scores.

There is a second major pitfall in using norm referenced tests to evaluate student progress. When this form of measurement is used, the results are typically reported in grade scores, percentiles or stanines. While there are some advantages to using each of these forms of reporting, all three have in common the limitation of unequal distances between specified intervals. For example, a year of improvement of reading where a student moves from a grade score of 1.0 to 2.0 most likely represents a substantially greater amount of improvement than an increase in grade score from 7.0 to 8.0. In almost all tests, the student who moved from 1.0 to 2.0 would have to show a significantly greater increase in raw score points when the pre and post tests are compared, than would the student moving from 7.0 to 8.0. Massad (1972) also discusses the way in which percentiles and stanines are similarly affected. Keeping records of the number of words learned, the specific skills mastered and the accuracy in orally reading some passages are all simple criterion referenced evaluations. This type of information should be kept and explained convincingly to administrators and the parents of the students. Millman (1970) and Airasian and Madaus (1972) outline ways in which criterion referenced measures can be used for reporting pupil progress. They suggest that checklists of instructional goals achieved by students be used to replace traditional grading procedures. The latter, Millman feels, have been devices for ranking, not rating students. Simple criterion

referenced measures can also be very effective in helping the pupil and the teacher to see the amount of progress being made.

Another legitimate procedure might be to randomly select a passage from material that is about to be used for instruction. Evaluate and record the student's oral reading, application of specific skills and his ability to answer questions asked to evaluate his understanding. This same passage can be readministered at the end of a teaching unit and the results compared. It might be possible to make statements like: When given a second grade level passage a student prior to instruction with materials in this chapter could: orally read it with 85% accuracy, made 8 word substitutions, refused to try two compound words (everything and anytime), could not state the main idea of the paragraph, and could answer 50% of the factual recall questions. At the conclusion of the chapter he read it with 92% accuracy, made three word substitutions, identified all the compound words, could still not state the main idea of the paragraph, and answered 90% of the factual recall questions. In addition to providing a wealth of evidence of progress, the analysis also offers suggestions for the next stage of instruction.

Some might object that the above procedure is inappropriate because the student was taught the material that he was later tested upon and that the practice effects of using it as a pre and post test may have artificially inflated the accuracy of the performance. From one point of view, it seems wasteful to spend time criticizing measurement techniques when effective improvement took place. A problem would arise only if the teacher spent an inordinate amount of time and effort when working with the randomly selected passage as compared with working with all other passages of the unit.

Teachers and reading specialists spend a great deal of time complaining about norm referenced testing. In most cases, however, they do not present alternative

data to support their contention that progress has been made that is not reflected in the norm referenced scores. Criterion referenced measures based on analysis and sampling of instructional materials offers a very reasonable means of accomplishing this.

There are some cautions in the use of criterion referenced measurements.

1. They cannot totally replace norm referenced evaluations. The latter are really better in answering some questions.
2. Criterion referenced measurement will probably fail if it is expected to emerge full-blown. While it might prove very satisfactory for three or four teachers to spend a year preparing one year's worth of objectives and test items, most clinical situations could hardly afford this time commitment.
3. There should be close correspondence between criterion referenced measures and instruction. With our present effectiveness in teacher preparation and our knowledge of the relative efficiency of strategies for teaching reading, it seems far better to set goals for programs and then devise measurement techniques which parallel these goals.
4. Data gathered during teaching should be considered valid indicators that criteria or goals have been met. If this is not accepted, an inordinate amount of time is spent on testing and insufficient time remains for teaching.
5. There is danger in dissecting reading into specific skills to a point where inefficient teaching results in the development of non-functional reading skills.

In spite of the cautions and difficulties criterion referenced measurement should be an essential activity, particularly in clinical reading programs. With the small number of students that are typically part of clinical instruction, individualization of instruction and goals should be possible. Criterion referenced measures are simply ways of determining whether goals have or have not been met.

Airasian, P. and Madaus, G. Criterion-referenced testing in the classroom. Measurement in Education, 1972, 3, 1-8.

Block, James H. Criterion-referenced measurement: Potential. The School Review, 1971, 79, 289-298.

Bleismer, E. Informal teacher testing in reading. The Reading Teacher, 1972, 26, 268-272.

Carver, R. Reading tests in 1970 versus 1980: Psychometric versus edumetric. The Reading Teacher, 1972, 26, 299-302.

Ebel, Robert L. Criterion-referenced measurement: Limitations. The School Review, 1971, 79, 282-298.

Glaser, R. Instructional technology and measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Hackett, Marie G. Criterion Reading, New York: Random House, Inc., 1970.

Hieronymous, A. Evaluation and reading: Perspective '72. The Reading Teacher, 1972, 26, 264-267.

Johnson, Marjorie S., and Kress, Roy A. Task analysis for criterion-referenced tests. The Reading Teacher, 1971, 24, 355-359.

Massad, C. Interpreting and using test norms. The Reading Teacher, 1972, 26, 286-292.

Millman, Jesson. Criterion-referenced measurement: An alternative. The Reading Teacher, 1972, 26, 278-281.

Pophem, W. James. The instructional objectives exchange: New support for criterion-referenced instruction. Phi Delta Kappan, 1970, 52, 174-175.

Pophem, W. James and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, Vol. 6, No. 1, 1-9.

Prescott, George A. Criterion-referenced test interpretation in reading. The Reading Teacher, 1971, 24, 347-354.

Proger, Barton B. and Mann, Lester. Criterion-referenced measurement: The world of gray vs black and white. Journal of Learning Disabilities, 1973, 6, 72-84.

Ransom, G. Criterion referenced tests - let the buyer beware. The Reading Teacher, 1972, 26, 282-285.

Simon, George B. Comments on "Implications of criterion-referenced measurement." Journal of Educational Measurement, 1969, 6, 259-260.